Detecting the stationarity of spatial dependence structure using spectral clustering

Véronique Maume-Deschamps, Pierre Ribereau, Manal Zeidan

Detecting the stationarity of spatial dependence structure using spectral clustering

Véronique MAUME-DESCHAMPS[1], Pierre RIBEREAU[1] and Manal ZEIDAN[1,2]

[1]Institut Camille Jordan, Univ Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5208, F-69622 Villeurbanne, France.
[2]Department of Operation Research and Intelligent techniques, University of Mosul, Mosul,Iraq.

Contributing authors: veronique.maume@univ-lyon1.fr; pierre.ribereau@univ-lyon1.fr; manal.zeidan@univ-lyon1.fr;

Abstract

Modeling extreme events requires the understanding of the spatial dependence structure in order to construct reliable statistical models. Assuming the stationarity of the dependence structure of the spatial process may not be reasonable, depending on the topology of the region under study for example. In environmental extreme events, different types of extremal dependencies could appear across the spatial domain. In this study, we present an adapted spectral clustering algorithm for spatial extremes by combining spectral clustering with extremal concurrence probability. This algorithm involves a heuristic method that can detect non stationarity in the dependence structure. In the case of a non-stationary dependence structure, the algorithm clusters the stations into $\boldsymbol{k}$ regions so that each region has a stationary dependence structure. To validate the proposed methodology, we tested it on different simulation cases using one or more max-stable models. The accuracy of the results encouraged us to apply it on two real data sets: rainfall data in the east coast of Australia and rainfall over France.

Keywords: Max-stable processes, Non-stationary dependence structures, Extremal concurrence probability, Spectral clustering

1 Introduction

Constructing a reliable statistical model for environmental extreme events, such as rainfall and temperature, is very important for understanding their behavior and accurately predicting their occurrence. Max-stable processes are natural models for spatial extremes, as they are natural extensions of the Extreme Value Theory (EVT) to spatial domains. They are powerful statistical models for extreme events in a continuous space and can assess the risk in areas that do not have stations. One basic assumption used in modeling is the stationarity of the dependence structure, but this assumption may be incorrect and can lead to the construction of meaningless models. In particular, if the data sets are taken from a large region or from regions with complex spatial features, it is plausible that the dependence structure will appear non stationary (Richards and Wadsworth (2021)). A non-stationary spatial dependence structure refers to the situation where the strength of the spatial dependence between extremes of a spatial process varies across the spatial domain.

In fact, dealing with non stationary spatial dependence structures is difficult in practice. Several approaches have been presented for modeling non-stationary dependence structures. For instance, Huser and Genton (2016) developed an approach that captures non-stationary patterns in spatial extremes using covariates. This method combines max-stable processes with a non-stationary correlation function. However, it requires prior knowledge of relevant covariates. Castro-Camilo and Huser (2020) developed a new methodology for modeling sub-asymptotic spatial extremes observed over large, heterogeneous regions using factor copula models. The proposed approach is able to capture complex non-stationary patterns and is well-suited for situations where the dependence strength weakens as events become more extreme. This methodology is flexible and efficient but it is computationally expensive. Richards and Wadsworth (2021) presented an approach for modeling nonstationary extremal dependence. They adapted deformation methods for spatial extremes by using least squares minimization of the difference between theoretical and empirical extremal dependence measures as a new objective function. Although this approach is effective, the estimation of the deformed space can be challenging. If the focus is on understanding the spatial patterns of extreme events, independent stationary dependence structures in different regions can be useful for modeling non-stationary dependence. This approach provides a simple and computationally efficient way to model spatial dependence in the data.

Recently, clustering was used to create regionalisations of the extreme events. Clustering is an unsupervised machine learning tool that is widely used in data analysis to identify subgroups with similar features. It has a wide range of applications in fields such as computer science, statistics, biology, and climate science.

In the context of spatial extremes, only a few studies have used clustering to partition an entire region into homogeneous sub-regions based on similarities in dependence structure. For instance, Bernard et al (2013) presented a novel clustering algorithm for maxima, using the F-madogram introduced

by Cooley et al (2006). By combining the F-madogram with a partitioning around medoids (PAM) algorithm, they were able to cluster the extremes based on dependence strength. The algorithm was applied to analyze rainfall patterns over France. Afterward, Bador et al (2015) applied this algorithm to large regions and different variables, analyzing the maxima of summer temperatures across Europe. Saunders et al (2021) demonstrated that, the PAM algorithm that presented by Bernard et al (2013) is sensitive to stations density. To address this issue, they proposed the use of hierarchical clustering with F-madogram. Then applied their proposed algorithm to rainfall stations in Australia and compared the resulting clusters to those obtained by the PAM algorithm.

Our main goal in this work is to investigate whether the spatial process under study has a stationary dependence structure or not, and if so, we aim to cluster the spatial process into k regional clusters, each with its own stationary dependence structure. To achieve this goal, we have adapted spectral clustering for spatial extremes by combining it with the extremal concurrence probability introduced by Dombry et al (2018). We also propose a heuristic method capable of detecting the stationarity of the dependence structure. The extremal concurrence probability for a max-stable process is the probability that the maximum value of the process occurs at two or more stations simultaneously. It is an important concept in the statistical modeling of extreme events, since the extremes exhibit concurrence, meaning that they have the same dependence structure. This motivated us to use it in conjunction with spectral clustering to identify regions with a stationary dependence structure. This combination of tools makes the proposed algorithm efficient in automatically determining the number of clusters and accurately clustering each station into its own group. We validated our method through a simulation study and then applied it to two sets of real data. The first dataset consists of rainfall data in the east coast of Australia, while the second dataset includes rainfall data over France provided by Météo-France.

The paper is organized as follows. Section 2 presents Max-stable processes. An overview of spectral clustering is provided in Section 3. Section 4 describes the adapted spectral clustering for spatial extremes. A simulation study is presented in Section 5. Section 6 applies the methodology to real data: rainfall over east coast of Australia and rainfall over France. Finally, Section 7 presents the discussion and conclusions of our study.

2 Max-stable processes

In this section, we will provide a brief overview of max-stable processes and define the extremal concurrence probability, which is a critical tool in our research.

2.1 Definition of Max-stable processes

Let $Z_1(s), Z_2(s) \cdots$ be a sequence of independent replications of a spatial process $\{Z(s), s \in \mathcal{S}\}, \mathcal{S} \subset \mathbb{R}^d, d \geq 1$. If there exist continuous functions $A_n(s) > 0$ and $B_n(s) \in \mathbb{R}$ such that

$$\frac{\max_{i=1,\cdots,n} Z_i(s) - B_n(s)}{A_n(s)} \overset{d}{=} X(s), s \in \mathcal{S}, n \to \infty, \tag{1}$$

is non-degenerate, then $\{X(s), s \in \mathcal{S}\}$ is a max stable process (see De Haan and Pereira (2006)). The univariate maxima $X(s)$ at any location s, follows a Generalized Extreme Value distribution (GEV), i.e, for all $x \in \mathbb{R}$,

$$\mathbb{P}(X(s) \leq x) = \exp[-(1 + \xi(s)\frac{x - \mu(s)}{\sigma(s)})^{-1/\xi(s)}], \tag{2}$$

where $\mu(s) \in \mathbb{R}$ is the location parameter, $\sigma(s) > 0$ is the scale parameter and $\xi(s) \in \mathbb{R}$ is the shape parameter. These parameters are possibly different from one location to another. Setting $\mu(s) = \sigma(s) = \xi(s) = 1$, leads to consider unit Fréchet distributions, i.e, $\mathbb{P}(X(s) \leq x) = \exp[-1/x], x > 0$, and $\{X(s), s \in \mathcal{S}\}$ is called a simple max-stable process (see Ribatet (2017) and Ribatet et al (2016)). De Haan (1984) provided the spectral representation for simple max-stable processes $\{X(s), s \in \mathcal{S}\}$ as follows:

$$X(s) = \max_{i \geq 1} \zeta_i Y_i(s), s \in \mathcal{S}, \mathcal{S} \subset \mathbb{R}^d, d \geq 1 \tag{3}$$

where $\{\zeta_i, i \geq 1\}$ is a Poisson point process on $(0, \infty)$ with intensity $\zeta^{-2} d\zeta$ and $Y_1(s), Y_2(s), \cdots$ denote a sequence of independent replications of a positive process $\{Y(s), s \in \mathcal{S}\}$ with $\mathbb{E}[Y(s)] = 1$ for all $s \in \mathcal{S}$.
Equation (3) may be written as follows:

$$X(s) = \max_{\varphi \in \Phi} \varphi(s), s \in \mathcal{S} \tag{4}$$

where $\Phi = \{\varphi_i(s) = \zeta_i Y_i(s) : s \in \mathcal{S}, i \geq 1\}$ is a Poisson point process on $\mathbb{C}_0$, the space of non-negative continuous functions on $\mathcal{S}$ (see Ribatet (2017)).
Let $\mathbf{S}$ be a set of m spatial locations : $\mathbf{S} = \{s_1, \cdots, s_m\} \subset \mathcal{S}$, then the multivariate maxima distribution is given by

$$\mathbb{P}\{X(s_1) \leq x_1, \cdots, X(s_m) \leq x_m\} = \exp\left\{-\mathbb{E}\left[\max_{j=1,\cdots,m} \frac{Y(s_j)}{x_j}\right]\right\} \tag{5}$$

where $\{Y(s), s \in \mathcal{S}\}$ is the process appearing in Equation (3). The exponent function

$$V_{\mathbf{S}}(x_1, \cdots, x_m) = \mathbb{E}\left[\max_{j=1,\cdots,m} \frac{Y(s_j)}{x_j}\right], \tag{6}$$

is called the exponent measure. It characterizes the dependence structure of $X(s_1), \cdots, X(s_m)$. Since the exponent measure is homogeneous of order -1, we can obtain a useful relation by setting $x_j = x$ for all $j = 1, \cdots, m$ such that $V_{\mathbf{S}}(1, \cdots, 1) = \theta_{\mathbf{S}}$ where $\theta_{\mathbf{S}}$ is the extremal coefficient that provides a summary of the dependence structure (see Schlather and Tawn (2003) and Smith (1990)). Particularly, when $\mathbf{S} = \{s_1, s_2\}$ the extremal coefficient satisfies $\theta_{\mathbf{S}} = V_{\mathbf{S}}(1, 1) \in [1, 2]$. The lower bound corresponds to the variables $X(s_1)$ and $X(s_2)$ which are completely dependent, while the upper bound corresponds to the case where they are independent.
Several models for max-stable processes have been presented based on this spectral representation, including the Brown-Resnick model (see Brown and Resnick (1977)), the Smith model (see Smith (1990)), the Schlather model (see Schlather (2002)), and the Extremal-t model (see Opitz (2013)).

2.2 Extremal concurrence probability

Other indices in order to measure the dependence between extremes exist in the literature. Dombry et al (2018) introduced the extremal concurrence probability for the analysis of extremal dependence, which was especially designed for max-stable processes. It has properties similar to the pairwise extremal coefficient, but it has the advantage of being a probability measure, which makes it more interpretable and axiomatic. The extremal concurrence probability focuses on the occurrence times of extremes, which means whether the record maxima occurs simultaneously, i.e., at the same time for all locations. It can be interpreted as the chance of a single extreme event affecting all the locations and being responsible for the record maximum.
It is based on the spectral representation of the max-stable processes. The idea behind this metric can be explained as follows.
Recall the spectral representation in Equation (4). We say that the extremes are concurrent at locations $s_1, \cdots, s_m \in \mathcal{S}$ if

$$X(s_j) = \varphi_\ell(s_j), j = 1, \cdots, m \tag{7}$$

for some $\ell \geq 1$. This means that the values of the process $\{X(s), s \in \mathcal{S}\}$ at locations $s_1, \cdots, s_m$ come from the same spectral function φ_ℓ.
The extremal concurrence probability is defined as

$$p_r(s_1, \cdots, s_m) = \mathbb{P}\{\text{for some } \ell \geq 1 : X(s_j) = \varphi_\ell(s_j),\ j = 1, \cdots, m\} \tag{8}$$

According to Theorem 3 in Dombry et al (2018), the bivariate extremal concurrence probability estimation coincides with Kendall's τ statistic:

$$\hat{p}_r(s_1, s_2) \equiv \tau = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \text{sign}\{X_i(s_1) - X_j(s_1)\}\text{sign}\{X_i(s_2) - X_j(s_2)\}, \tag{9}$$

where $\{X_i(s), s \in \mathcal{S}, i = 1, \cdots, n\}$ are n independent copies of $\{X(s), s \in \mathcal{S}\}$. The bivariate extremal concurrence probability for max-stable processes satisfies $p_r(s_i, s_j) = 0$ if and only if $X(s_i)$ and $X(s_j)$ are independent, and $p_r(s_i, s_j) = 1$ if and only if $X(s_i)$ and $X(s_j)$ are almost surely equal. These properties were stated and proved in Proposition 1 of Dombry et al (2018).

3 Spectral clustering : an overview

Spectral clustering is a technique used in machine learning and data analysis for clustering data points into groups based on the similarity between them. It is based on the concept of spectral graph theory, which is the study of the properties of graphs using linear algebra.
Spectral clustering has several advantages, as it can handle high-dimensional data, which is often a limitation for other clustering algorithms. This is done by reducing the high-dimensional data to a lower-dimensional space using eigenvalue decomposition. Furthermore, it can handle different kinds of similarity measures, which makes it flexible and adaptable to different types of data. Also, it does not make any assumptions on the shape or size of clusters.
Spectral clustering considers the dataset as a graph, where each data point $x_i, i = 1, \cdots, n$ represents a vertex in an undirected weighted graph. An undirected graph $G = (V, E, S)$ is generally defined by a set of vertices $V = \{v_1, v_2, \cdots, v_n\}$, a set of edges $E = \{(v_i, v_j)|v_i, v_j \in V\}$ between these vertices, and a similarity matrix S. An element $s_{ij} \in S$ represents the amount of similarity between the vertices v_i,v_j and the weight that will be assigned to each edge. It is important to note that since the graph is undirected, the similarity matrix should be symmetric. If $s_{ij} = 0$, this means that there is no edge between the vertices v_i,v_j. Each vertex v_i in the graph has a degree d_i:

$$d_i = \sum_{j=1}^{n} s_{ij}. \tag{10}$$

The degrees $d_1, \cdots, d_n$ represent the elements of a diagonal matrix called the degree matrix of the graph D.
Spectral clustering aims to separate the main graph G into sub-graphs so that the weights of the edges between these sub-graphs are small, indicating dissimilarity between the clusters, while the weights of the edges connecting nodes within each sub-graph are relatively high, indicating similarity within the clusters.

3.1 Steps of spectral clustering algorithm

In general, any spectral clustering algorithm involves the following three steps.

1. **Pre-processing**
   Construct the similarity matrix S from the dataset using a measure that takes into account the aim of clustering, and then construct the similarity

graph. There are different ways to do this depending on the pairwise similarity s_{ij}. The aim is to model the neighborhood relation among the data points $x_1, \cdots, x_n$. These ways are summarized as follows:

- ε **-neighborhood graph:** In this graph, the vertices v_i, v_j will be connected by an edge if they are similar enough, i.e if $s_{ij} > \varepsilon$, ε is a pre-defined non-negative real number. Usually, this graph is considered as an unweighted graph.
- k**-nearest neighbor graphs:** In this graph, the distance between each pair of vertices is computed using the Euclidean distance. Then, the vertices v_i, v_j are connected by an edge if v_j is among the k nearest neighbors of v_i or vice versa, and the edge is weighted by the similarity s_{ij}. The neighborhood relationship among data points is controlled by a pre-defined integer number k.
- **The fully connected graph:** In this graph, each vertex is connected to all other vertices by edges, and these edges are weighted by the similarities s_{ij}. This type of graph is useful only if the similarity function can model the neighborhood relation among the data points. The commonly used similarity function is the Gaussian similarity function, which is defined as $s_{ij} = \exp(-\|x_i - x_j\|^2/2\sigma^2)$, where the neighborhood relation is controlled by σ.

For further information on similarity graphs, we refer to Von Luxburg (2007) and Parodi (2012).

2. **Spectral representation**

Compute the Laplacian matrix of the graph, which is an essential tool to identify clusters in the data using spectral clustering. It is a matrix that characterizes the connectivity of a graph. It captures the relationships between the nodes, and can be used to identify the nodes that are most closely connected to each other. There are two different definitions for this matrix, depending on the degree matrix D and the similarity matrix S of the graph, as follows.

(a) Unnormalized Laplacian matrix L: $L = D - S$.

(b) Normalized Laplacian matrix L^{sym}: $L^{sym} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$.

The choice of Laplacian matrix type to use with spectral clustering depends on the application and the problem to be solved. Spectral clustering is often used to optimize two objective functions: Ratio Cut (Rcut) and Normalized Cut (Ncut). Both of these objective functions measure the quality of the partition of a graph into clusters. Let C_i be a subset of vertices i.e $C_i \subset V, i = 1, \cdots, k$ and its complement $\overline{C}_i := V \backslash C_i$, the Ratio Cut function (Rcut) (Hagen and Kahng (1992)) is defined as:

$$\text{Rcut}(C_1, \cdots, C_k) = \sum_{i=1}^{k} \frac{\text{cut}(C_i, \overline{C}_i)}{|C_i|}. \tag{11}$$

Where

$$\text{cut}(C, \overline{C}) := \sum_{i \in C, j \in \overline{C}} s_{ij}$$

$$|C_i| := \text{number of vertices in } C_i$$

In this function, the size of a subset C_i is measured by its number of vertices. Using the unnormalized Laplacian matrix L with spectral clustering leads to minimizing the Ratio Cut function.

In contrast, the Normalized cut function(Ncut)(Shi and Malik (2000)) is defined as:

$$\text{Ncut}(C_1, \cdots, C_k) = \sum_{i=1}^{k} \frac{\text{cut}(C_i, \overline{C}_i)}{\text{vol}(C_i)}. \tag{12}$$

where

$$\text{vol}(C) := \sum_{i \in C} d_i$$

In the Normalized cut function, the size of a subset C_i is measured by the weights of its edges. Using the normalized Laplacian matrix L^{sym} with spectral clustering leads to minimizing the Normalized cut function. For more details see Von Luxburg (2007).

The matrices L and L^{sym} have some important properties: they are symmetric and positive semi-definite matrices; the n eigenvalues $\lambda_1, \cdots, \lambda_n$ of these matrices are non negative real-valued, so $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$; the multiplicity k of the value 0 as an eigenvalue of these matrices is equal to the number of connected components $C_1, \cdots, C_k$ in the graph. (for more details, see Mohar et al (1991), Mohar (1997) and Chung (1997)).

The eigenvalues of the graph Laplacian matrix and its associated eigenvectors are computed. Then, the eigenvectors are used to constitute a low-dimensional representation of the data, where the clusters are more separated. Typically, the k eigenvectors corresponding to the k smallest eigenvalues are used to construct a k-dimensional representation of the data, as they capture the structure of the graph and important features of the data (see Wierzchoń and Kłopotek (2018)). Reducing dimension can reveal hidden patterns in the data that may be difficult to distinguish in higher dimension.

3. **clustering**
   Apply the k-means clustering algorithm to the low-dimensional representation to group the data points into k clusters.

3.2 A heuristic method to determine the number of clusters *k*

In spectral clustering, a specific heuristic method has been proposed for choosing the number of clusters k. This method relies on the gap between two consecutive eigenvalues, with the number of clusters determined by the value of k that maximizes the eigengap δ_k: $\delta_k = \mid \lambda_{k+1} - \lambda_k \mid$, $k \geq 2$ (see Von Luxburg (2007)). This method is effective in determining the number of clusters when

the dataset is well separated. However, choosing k greater than or equal to 2 leads to clustering the dataset into at least two groups, which may not be appropriate if the goal is to verify whether the dataset can be considered a single group (more precisely, if the spatial process has a stationary dependence structure). To address this issue, we propose to add another heuristic methodology as the first step to check if the data can be considered as a single group before making clusters.

The underlying idea behind this heuristic methodology comes from the fact that the second smallest eigenvalue λ_2 of the Laplacian matrix corresponds to the algebraic connectivity, also known as the Fiedler value. This value reflects how well the overall graph is connected (Fiedler (1973)) and provides information about the intensity of the connections between the nodes of the graph.

If λ_2 is small, it suggests that the graph is nearly disconnected, and vice versa (see Wierzchoń and Kłopotek (2018)). In other words, when the graph is well connected, λ_2 is large and far from the first eigenvalue. Since we have $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ for a graph Laplacian matrix, if we examine the first ten eigenvalues in the set, we will observe that λ_1 is the only outlier value. This scenario indicates that the graph is well-connected, and the data is a single group. Conversely, if the graph can be partitioned into sub-graphs, we can use the largest eigengap δ_k to determine the number of clusters k. A simulation study in Section 5 validates this method based on the largest value of δ_k. The steps of this heuristic methodology are described in Algorithm 1.

Algorithm 1 Heuristic method to determine the number of clusters k

Require: Vector of eigenvalues λ.

Ensure: Number of clusters k.

1: Find the outliers value in the eigenvalues set $(\lambda_1, \cdots, \lambda_{10})$.
2: If λ_1 is the only outlier value, then $k = 1$. Else, go to step 3.
3: Calculate the eigengap δ_k: $\delta_k = \mid \lambda_{k+1} - \lambda_k \mid$, $k \geq 2$.
4: k corresponds to the largest value of δ_k.

4 Adapting spectral clustering for spatial extremes

Let $X_i(s_j), s_j \in \mathcal{S}$, $\mathcal{S} \subset \mathbb{R}^d$, $d = 2$, $i = 1, \cdots, n$ be a sequence of n independent and identically distributed max stable processes at different locations $s_j, j = 1, 2, \cdots, m$. In order to apply spectral clustering in a spatial extreme context, locations $s_1, \cdots, s_m$ are considered as vertices in a fully connected graph. Each vertex is connected to all others by edges, and the weights (similarity values) of these edges represent the dependence strength among the locations. For viewing purposes, let us assume that the max-stable process consists of 15 locations. The fully connected graph is shown in Figure 1.

Selecting an appropriate metric to construct the similarity matrix is essential

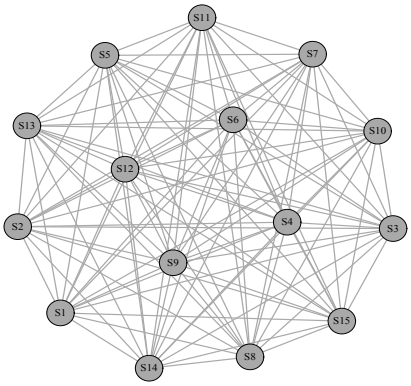

Fig. 1 Fully connected graph with 15 vertices. Each vertex represents a location in the max-stable process

in the spectral clustering algorithm, especially when using a fully connected graph. It is important to choose a spatial dependence measure that can accurately model the neighborhood relations among the locations. In this study, we used the extremal concurrence probability, as introduced by Dombry et al (2018) (see Section 2.1). The similarity matrix represents the pairwise extremal concurrence probability matrix, denoted by $CP \in \mathbb{R}^{m\times m}$, where m is the number of locations. For a pair $(s, s') \in \mathcal{S} \times \mathcal{S}$, the element of the matrix CP is given by:

$$\hat{p}_r(s, s') = \frac{2}{n(n-1)} \sum_{1\leq i<j\leq n} \text{sign}\{X_i(s) - X_j(s)\}\text{sign}\{X_i(s') - X_j(s')\} \quad (13)$$

After constructing the similarity matrix CP according to Equation (13), it is used to compute the graph Laplacian matrix. Using the normalized graph Laplacian matrix L^{sym} helps to achieve our goal of making the size of the resulting clusters dependent on the strength of the dependence structure (i.e., the weights of the graph edges). The spectrum λ of L^{sym} is then computed and used as input in Algorithm 1 to determine the number of clusters k. If $k \geq 2$, the eigenvectors $q_1, \cdots, q_k$ are used to constitute a k-dimensional representation for the data. This is done by representing these eigenvectors as columns of an $m \times k$ matrix denoted Q. Each row in Q represents a location s_j: $s_j \rightarrow (q_{j,1}, \cdots, q_{j,k})$, $j = 1, \cdots, m$, this is called spectral mapping (see Wierzchoń and Kłopotek (2018)). Normalizing each row of Q to norm 1 results in a matrix denoted $Y \in \mathbb{R}^{m\times k}$. According to Ng et al (2001), this last step improves the performance of the clustering algorithm. Instead of using k-means, which is usually used at this step, we used a Gaussian Mixture Model (GMM) to cluster the rows of Y. GMM clusters the datapoints based on probability distribution, considering that the datapoints come from a Gaussian mixture. Each cluster has a Gaussian distribution model with parameters mean and covariance. Taking the covariance into account makes GMM more robust

than k-means, which depends only on the cluster mean. For more details about GMM, see for example Bouveyron et al (2019). We summarize these steps in Algorithm 2.

After multiple attempts, the extremal concurrence probability has demon-

Algorithm 2 Proposed spectral clustering

Require: The similarity matrix $CP \in \mathbb{R}^{m\times m}$, constructed according to Equation (13) .

Ensure: Clusters $\{C_1, \cdots, C_k\}$.

1: Compute the normalized Laplacian matrix $L^{sym} = D^{-\frac{1}{2}}(D - CP)D^{-\frac{1}{2}}$.
2: Compute the spectrum of L^{sym} and use Algorithm 1 in order to determine k. If $k = 1$, the algorithm stops. Else, go to step 3.
3: Compute the k smallest eigenvectors $q_1, q_2, \cdots, q_k$ of L^{sym}, and arrange these vectors in columns to be a matrix Q, where $Q \in \mathbb{R}^{m\times k}$.
4: Normalize the rows of Q to norm 1, resulting the matrix $Y \in \mathbb{R}^{m\times k}$: $Y_{jl} = Q_{jl}/(\sum_l Q_{jl}^2)^{\frac{1}{2}}$, $j = 1, \cdots, m$, $l = 1, \cdots, k$.
5: Consider each row of Y as a point in $\mathbb{R}^k$ and implement Gaussian Mixture Model (GMM) to cluster them into k clusters.
6: Assign the location s_j to cluster l if and only if row j of the matrix Y is assigned to cluster l.

strated its ability to detect different types of spatial dependence when used as a similarity matrix in spectral clustering, compared to other extremal dependence measures. In order to illustrate this point, we have simulated two spatial processes with 15 locations generated randomly and uniformly in $[0, 1]^2$. The number of observations was set to 1000. In the first simulation, we randomly selected 5 locations and used them to simulate a Brown-Resnick model, while the remaining 10 locations were used to simulate a Schlather model. The parameters for both models were arbitrarily chosen. In the second simulation, we randomly chose six locations (five and four, respectively) to simulate a Brown-Resnick max-stable model with parameters arbitrarily chosen. Then we attempted to use different spatial dependence measures as a similarity matrix, such as the extremal concurrence probability (Dombry et al (2018)), extremal coefficient (Schlather and Tawn (2003)), and F-madogram (Cooley et al (2006)). We have computed their normalized Laplacian matrix and plotted their eigenvalues against eigenvectors in order to compare the ability of these measures to detect different types of dependence structures in the data. In the case of the first simulation, Figure 2 shows a plot of eigenvalues against eigenvectors for the normalized Laplacian matrix of the pairwise extremal concurrence probability matrix, pairwise extremal coefficient matrix, and pairwise F-madogram matrix, respectively. The left panel in Figure 2 shows that λ_2 has a low value, suggesting that the graph is nearly disconnected. The maximum eigengap lies between λ_3 and λ_2, indicating that the data consists of two groups, each with a different type of dependence structure. This is consistent

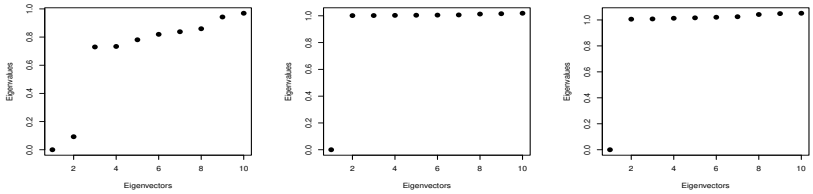

Fig. 2 The eigenvalues against eigenvectors for the normalized Laplacian matrix of the pairwise extremal concurrence probability matrix, pairwise extremal coefficient matrix, and pairwise F-madogram matrix in the left, middle, and right panels, respectively, for the first simulation.

with the data of the first simulation. Figure 3 shows the pairwise extremal concurrence probability matrix for this simulations data before and after using the proposed spectral clustering to cluster the locations. On the other hand, the middle and right panels in Figure 2 reveal that λ_2 has a high value, indicating a well-connected graph and only one type of dependence structure, which contradicts the expected result with two groups. In the case of the sec-

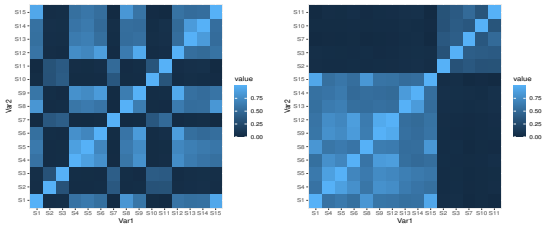

Fig. 3 The left panel shows the pairwise extremal concurrence probability matrix for the data of the first simulation before clustering, while the right panel shows the matrix after using the proposed spectral clustering to cluster the locations.

ond simulation, Figure 4 shows a plot of eigenvalues against eigenvectors for the normalized Laplacian matrix of the pairwise extremal concurrence probability matrix, pairwise extremal coefficient matrix, and pairwise F-madogram matrix, respectively. In the left panel in Figure 4, λ_2 has a low value, and the maximum eigengap is between λ_4 and λ_3. This indicates that the data consists of three groups, each with a different type of dependence structure, which is consistent with the data of the second simulation. Figure 5 shows the pairwise extremal concurrence probability matrix for this simulation data before and after using the proposed spectral clustering to cluster the locations. While the middle and right panels in Figure 4 show that the data has only one type of dependence structure, which contradicts the expected result with three groups. This indicates that the extremal coefficient and F-madogram measures are not appropriate to construct the similarity matrix, since they cannot detect different types of dependence structures in the data.

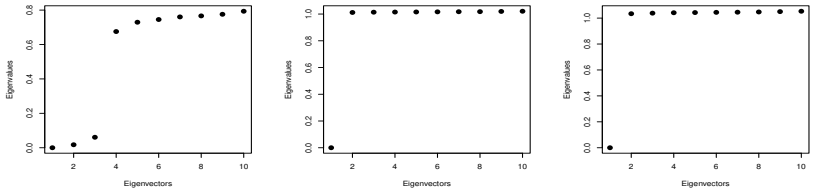

Fig. 4 The eigenvalues against eigenvectors for the normalized Laplacian matrix of the pairwise extremal concurrence probability matrix, pairwise extremal coefficient matrix, and pairwise F-madogram matrix in the left, middle, and right panels, respectively, for the second simulation.

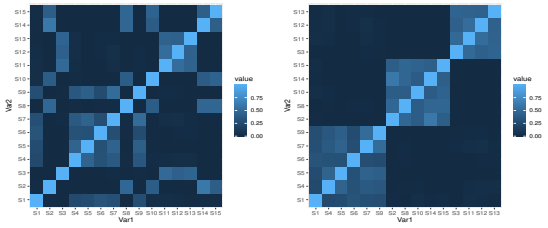

Fig. 5 The left panel shows the pairwise extremal concurrence probability matrix for the data of the second simulation before clustering, while the right panel shows the matrix after using the proposed spectral clustering to cluster the locations.

5 Simulation study

In order to assess the accuracy of our algorithm, we tested it on three simulation cases. In each case, we implemented the algorithm on several spatial processes simulated from one or more max-stable models: Smith, Schlather, Brown-Resnick and Extremal-t, with parameters chosen randomly. Furthermore, the correlation functions for Schlather and Extremal-t models were chosen randomly from one of the following correlation functions: Cauchy, powered exponential and Whittle-Matérn. The number of observations was fixed to 1000 at each location $s_j, j = 1, 2, \cdots, m$. To avoid any effect of the number of locations m on the algorithm performance, m was chosen randomly for each simulation, uniformly between 30 and 100 locations. The m locations were generated randomly and uniformly in $[0, 1]^2$. The three simulation cases are described in detail below.

- Case 1: stationary dependence structure
  We simulated 100 spatial processes. In each simulation, we randomly selected one of the max-stable models and simulated it on the m locations.
- Case 2: non-stationary dependence structure / different models
  We considered three sub-cases where k equals 2, 3, and 4 respectively. For each sub-case, we simulated 100 spatial processes. In each simulation, we randomly partitioned the m locations into k groups, so that $\sum_{\ell=1}^{k} m_\ell = m$.

We then randomly selected k different models from the max-stable models and used m_ℓ locations to simulate model ℓ.

- Case 3: non-stationary dependence structure / one model with different parameters
  We considered three sub-cases where k equals 2, 3, and 4 respectively. For each sub-case, we simulated 100 spatial processes. In each simulation, we randomly partitioned the m locations into k groups, so that $\sum_{\ell=1}^{k} m_\ell = m$. We then randomly selected one of the max-stable models and used m_ℓ locations to simulate it, with the parameters of model ℓ chosen randomly and differing from the parameters of other clusters.

We assess the accuracy of the proposed algorithm using two evaluation measures. The first one is the accuracy of the heuristic method for determining the number of clusters, which we will denote as $\mathcal{A}_k$ for simplicity:

$$\mathcal{A}_k = \frac{\sum_{t \subset T} \hat{k}_t}{T} \tag{14}$$

where T is the total number of spatial processes tested and $\hat{k}_t = 1$ if the estimation of the number of cluster is correct, while $\hat{k}_t = 0$ otherwise.
The second one is the accuracy of our proposed spectral clustering algorithm in correctly grouping the simulated locations from the same model into the same cluster. To evaluate this, we used the clustering purity measure (Schütze et al (2008)). It calculates the ability of a clustering method to recover known groups. Clustering purity is applicable even when the number of clusters k is different from the number of known groups. It is computed by assigning a label to each cluster based on the most frequent group in it, and then sum the number of correct group labels in each cluster and divide it by the total number of data points. Depending on the specific clustering issue at hand, the formula for the purity measure is:

$$Purity(\mathcal{C}, \mathcal{P}) = \frac{1}{m} \sum_{\ell=1}^{k} \max_{i=1,\cdots,g} |C_\ell \cap P_i| \tag{15}$$

Where $\mathcal{C} = \{C_1, \cdots, C_k\}$ is the set of identified clusters by spectral clustering, $\mathcal{P} = \{P_1, \cdots, P_g\}$ is the set of simulated groups, $|C_\ell \cap P_i|$ is the number of locations of cluster ℓ being in group i and m is the total number of locations. Clustering purity is a real number in [0,1]. A higher value of purity indicates better clustering performance, meaning that the algorithm can accurately identify clusters that correspond to the true groups of locations.
We computed the purity of the proposed spectral clustering algorithm using GMM and k-means in step 5 of the Algorithm 2. Since we applied the clustering method to 100 different spatial processes for each sub-case of the non-stationary dependence structure, we took the average purity. We computed both evaluation measures, $\mathcal{A}_k$ and average purity, for all simulation cases considered and presented the results in Table 1. Note that, for simplicity, we

denote the average purity of spectral clustering with GMM and k-means by P-SC(GMM) and P-SC(k-means), respectively.

Table 1 The evaluation measures for the proposed spectral clustering within all simulation cases

	Stationary [1]	Non-stationary [2]			Non-stationary [3]		
Evaluation measures	-	$k=2$	$k=3$	$k=4$	$k=2$	$k=3$	$k=4$
$\mathcal{A}_k$	1	1	1	0.99	1	1	1
P-SC(GMM)	-	1	1	0.9993	0.9994	1	1
P-SC(k-means)	-	1	0.9495	0.9481	0.9994	0.9386	0.9082

[1]Simulation case 1

[2]Simulation case 2

[3]Simulation case 3

Regarding the performance of the proposed heuristic method (first row of Table 1), it appears to be accurate in detecting whether the spatial process has a stationary dependence structure or not, as well as in determining the correct number of clusters, $\hat{k}$, in the non-stationary cases.
The second row of Table 1 shows that the proposed spectral clustering algorithm has a high accuracy in clustering the locations according to the model from which they were simulated. This includes the two simulation cases of non-stationary dependence structure with all tested numbers of clusters.
One can also note that the accuracy of spectral clustering with GMM and k-means is the same when the number of clusters equals 2, while for a number of clusters equal to 3 and 4, the spectral clustering with GMM is more accurate. In fact, the accuracy of spectral clustering with k-means decreases as the number of clusters increases, as noted in row three of Table 1.
Since the results of the simulations appear satisfactory, we can use this technique to detect stationary extreme areas for precipitation in East Australia and France.

6 Application on real data

This section is devoted to two real data applications: one on rainfall in Australia's east coast, and the other on rainfall in France.

6.1 Rainfall over east coast of Australia

We will begin with a brief description of the data, followed by the application of our clustering method and a discussion of the results.

6.1.1 Description of the data

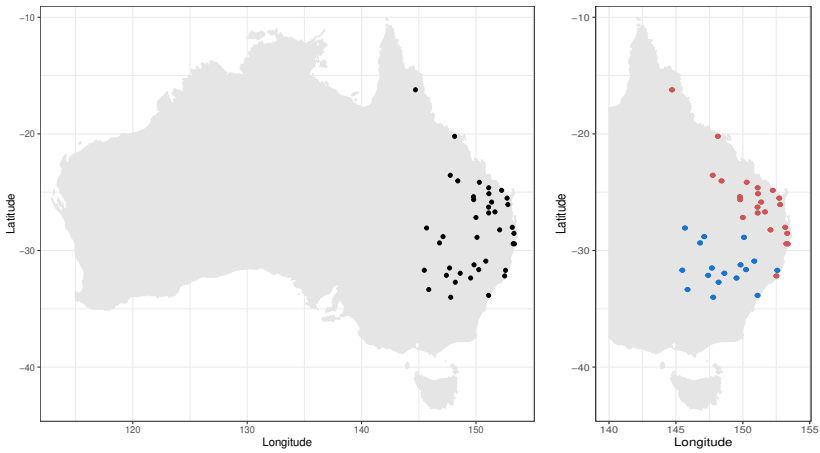

Fig. 6 Geographic locations of 40 stations on the east coast of Australia in the left panel, and proposed spectral clustering in the right panel.

This data represents the daily rainfall totals (in millimeters) measured over a 24-hour period at 40 stations on the east coast of Australia during the winter season (April to September) from 1972 to 2019, resulting in a total of 183 * 48 = 8,784 observations at each station. The altitude of these stations ranges from 2 to 540 meters. The geographic locations of the 40 stations are illustrated in the left panel of Figure 6. More information about this data can be found in references such as Ahmed et al (2022), Bacro et al (2016), Ahmed et al (2017), and Abu-Awwad et al (2020). The data is freely available on the website http://www.bom.gov.au.

6.1.2 Detecting the stationarity of Australia rainfall dependence structure

We will apply our proposed spectral clustering algorithm, described in Section 3 Algorithm 2, to detect the stationarity of rainfall dependence structure. To demonstrate the effect of the block size on the stationarity of the dependence structure, we will use different sizes of block. We will use the same sizes as Ahmed et al (2022). Specifically, we will test block sizes of 183 days, 30 days, 15 days, 10 days, 5 days, 3 days, and 1 day. Table 2 shows the results for detecting the stationarity of the dependence structure for Australia's rainfall data for each block size. We observe that the block size has an impact on the stationarity of the dependence structure. The rainfall dependence structure is identified as stationary when the block size is greater than or equal to 15 days. By using these block sizes, we can model the entire spatial data with a single model.

The rainfall dependence structure is non-stationary when the block size is less than or equal to 10 days. The heuristic method determines that the number of clusters is equal to 2 for all these block sizes. The spectral clustering algorithm clustered the 40 locations to two clusters each with stationary

Table 2 The results of detecting the stationarity of the dependence structure for Australia's rainfall data using different block sizes.

Block size	Spectral clustering implementation outputs		
	Stationary dependence structure	Non-stationary dependence structure	No. of clusters
183 days	✓	-	-
30 days	✓	-	-
15 days	✓	-	-
10 days	-	✓	$k = 2$
5 days	-	✓	$k = 2$
3 days	-	✓	$k = 2$
1 day	-	✓	$k = 2$

dependence structure. These clusters are the same for each of these block sizes and illustrated in the right panel of Figure 6. So, it is suitable to use two models when modeling this data with these block sizes. One model for the North and one for the South.

6.2 Rainfall over France

This subsection is devoted to the study of rainfall data in France.

6.2.1 Description of the data

This data is provided by Météo-France and represents the hourly precipitation recorded at 80 French monitoring stations. The data was measured during the fall season (September, October and November) over the period 1993 - 2021. Each station has 91 * 29 = 2639 observations. The geographic locations of these stations were chosen to cover all the French metropolitan regions. Figure 7 illustrates the geographic locations of the 80 stations. This data was studied by Bernard et al (2013) during the period 1993 - 2011.

6.2.2 Detecting the stationarity of France rainfall dependence structure

In order to detect the stationarity of the rainfall dependence structure, we implemented the proposed spectral clustering on the data. Additionally, we studied the effect of the block size on the stationarity of the dependence structure. The block sizes we considered are annual, monthly, 2 weeks, and weekly. The results obtained with the spectral clustering on this data for each block size are shown in Table 3.

It is clear that for both annual and monthly block sizes, the rainfall has a stationary dependence structure, while non-stationarity in the dependence structure appears when the block size is less than or equal to two weeks.

The clustering of stations obtained by spectral clustering for each block size

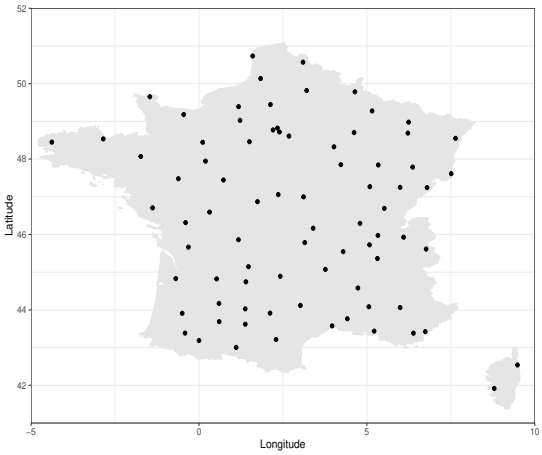

Fig. 7 Geographic locations of 80 stations over France region

Table 3 Detecting the stationarity of the dependence structure for France rainfall data during the period (1993 - 2021) with different sizes of block maxima.

Block size	Spectral clustering implementation outputs		
	Stationary dependence structure	Non-stationary dependence structure	No. of clusters
Annual	✓	-	-
Monthly	✓	-	-
2 Weeks	-	✓	$k = 2$
Weekly	-	✓	$k = 2$

leading to non-stationary dependence structure is illustrated in Figure 8, where the left panel is related to block size equal to two weeks and the right panel is related to the weekly block size.

We will begin our discussion with the weekly block size since it was studied by Bernard et al (2013). Implementing spectral clustering on the data with this block size shows that the rainfall has a non-stationary dependence structure. The spectral clustering divides France into two regional areas, north and south, along the Loire valley line. Each of these regions has a different dependence structure, as explained in the right panel of Figure 8. This can be interpreted easily. The extreme rainfall in the north of France is produced by disturbances from the Atlantic, which affect Brittany, Paris and other areas in the north of France. In contrast, the extreme rainfall in the south of France is caused by the Mediterranean sea, which affects the coastal areas, particularly Cévennes and the Montagne Noire. The results are similar to those obtained by Bernard et al (2013), where the selection criterion for the number of clusters indicated that $k = 2$. The locations clustering result is relatively close to Bernard et al (2013), where France was divided into north and south regions.

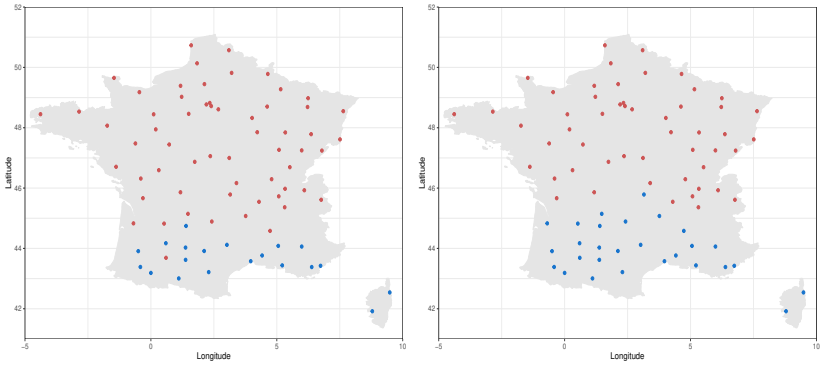

Fig. 8 The results of the proposed spectral clustering for the France rainfall data over the period of 1993-2021 are presented in both the left and right panels, with a block size equal to two weeks and the weekly block size, respectively.

This indicates that the behavior of the data has not changed since 2011.
The left panel of Figure 8 shows the stations clustering result when the size of block equals two weeks. It is clear that the south region is smaller than the ones in the weekly block size. This change in the size of clusters is due to the fact that the size of block had become larger compared to the weekly block size. This indicates that the larger size of block, the nearer to stationary in dependence structures.

7 Discussion and Conclusion

The stationarity of the dependence structure is an essential matter in modeling environmental extreme events. In most studies, it is assumed that the dependence structure is stationary. However, this may be incorrect, especially in large regions and regions with complex geographical or climatic patterns. Therefore, finding a method that can detect regions with a stationary dependence structure is useful. In this study, we combined spectral clustering with extremal concurrence probability to create a simple clustering method for max-stable processes. Additionally, we proposed a heuristic method to determine whether the dependence structure of the data are stationary or not.
We validated the proposed spectral clustering algorithm through a simulation study. Then we studied two environmental data sets. The first one is the daily rainfall data over east coast of Australia. We found that this data has a stationary dependence structure when the block size is larger than or equal to 15 days. The data has the same two regional clusters when the block size is smaller than or equal to 10 days. The second data set is the hourly precipitation over France. we found that stationarity appeared for large block sizes (monthly and annual), while non-stationary dependence structures are plausible for block sizes less than or equal to two weeks. The regional clusters are not the same for all these block sizes. Therefore, we conclude that the size of

the block affects the stationarity of the dependence structures and can result in changes in regional clusters. Thus, the difference in the dependence structure for small and large block sizes must be considered when modeling. For instance, different models can be used for different block sizes. A stationary model could be used for large block sizes, while for small block sizes, one could either identify independent stationary regions through clustering or use a non-stationary model.
Finally, despite the simplicity of the proposed algorithm, it is powerful. As a future direction of this study, one can study other variables like temperature. Another direction is to test the efficiency of this algorithm when applied to a very large region, such as the whole of Australia or the continent of Europe.

Acknowledgments We acknowledge partial support from the PAUSE program, which is operated by the Collège de France. Also, we would like to thank the French meteorological service (Météo-France) for providing us with the data.

References

Abu-Awwad A, Maume-Deschamps V, Ribereau P (2020) Fitting spatial max-mixture processes with unknown extremal dependence class: an exploratory analysis tool. Test 29(2):479–522. https://doi.org/10.1007/s11749-019-00663-5

Ahmed M, Maume-Deschamps V, Ribereau P, et al (2017) A semi-parametric estimation for max-mixture spatial processes. arXiv preprint arXiv:171008120

Ahmed M, Maume-Deschamps V, Ribereau P (2022) Recognizing a spatial extreme dependence structure: A deep learning approach. Environmetrics 33(4):e2714. https://doi.org/10.1002/env.2714

Bacro JN, Gaetan C, Toulemonde G (2016) A flexible dependence model for spatial extremes. Journal of Statistical Planning and Inference 172:36–52. https://doi.org/10.1016/j.jspi.2015.12.002

Bador M, Naveau P, Gilleland E, et al (2015) Spatial clustering of summer temperature maxima from the cnrm-cm5 climate model ensembles & e-obs over europe. Weather and climate extremes 9:17–24. https://doi.org/10.1016/j.wace.2015.05.003

Bernard E, Naveau P, Vrac M, et al (2013) Clustering of maxima: Spatial dependencies among heavy rainfall in france. Journal of climate 26(20):7929–7937. https://doi.org/10.1175/JCLI-D-12-00836.1

Bouveyron C, Celeux G, Murphy TB, et al (2019) Model-based clustering and classification for data science: with applications in R, vol 50. Cambridge

University Press

Brown BM, Resnick SI (1977) Extreme values of independent stochastic processes. Journal of Applied Probability 14(4):732–739. https://doi.org/10.2307/3213346

Castro-Camilo D, Huser R (2020) Local likelihood estimation of complex tail dependence structures, applied to us precipitation extremes. Journal of the American Statistical Association 115(531):1037–1054. https://doi.org/10.1080/01621459.2019.1647842

Chung FR (1997) Spectral graph theory, vol 92. American Mathematical Soc.

Cooley D, Naveau P, Poncet P (2006) Variograms for spatial max-stable random fields, Springer New York, New York, NY, pp 373–390. https://doi.org/10.1007/0-387-36062-X_17

De Haan L (1984) A spectral representation for max-stable processes. The annals of probability 12(4):1194–1204

De Haan L, Pereira TT (2006) Spatial extremes: Models for the stationary case. The annals of statistics 34(1):146–168. https://doi.org/10.1214/009053605000000886

Dombry C, Ribatet M, Stoev S (2018) Probabilities of concurrent extremes. Journal of the American Statistical Association 113(524):1565–1582. https://doi.org/10.1080/01621459.2017.1356318

Fiedler M (1973) Algebraic connectivity of graphs. Czechoslovak mathematical journal 23(2):298–305

Hagen L, Kahng AB (1992) New spectral methods for ratio cut partitioning and clustering. IEEE transactions on computer-aided design of integrated circuits and systems 11(9):1074–1085. https://doi.org/10.1109/43.159993

Huser R, Genton MG (2016) Non-stationary dependence structures for spatial extremes. Journal of agricultural, biological, and environmental statistics 21(3):470–491. https://doi.org/10.1007/s13253-016-0247-4

Mohar B (1997) Some applications of Laplace eigenvalues of graphs, Springer Netherlands, Dordrecht, pp 225–275. https://doi.org/10.1007/978-94-015-8937-6_6

Mohar B, Alavi Y, Chartrand G, et al (1991) The laplacian spectrum of graphs. Graph theory, combinatorics, and applications 2(871-898):12

Ng A, Jordan M, Weiss Y (2001) On spectral clustering: Analysis and an algorithm. Advances in neural information processing systems 14

Opitz T (2013) Extremal t processes: Elliptical domain of attraction and a spectral representation. Journal of Multivariate Analysis 122:409–413. https://doi.org/10.1016/j.jmva.2013.08.008

Parodi P (2012) Computational intelligence with applications to general insurance: a review: I–the role of statistical learning. Annals of Actuarial Science 6(2):307–343. https://doi.org/10.1017/S1748499512000036

Ribatet M (2017) Modelling spatial extremes using max-stable processes

Ribatet M, Dombry C, Oesting M (2016) Spatial extremes and max-stable processes. Extreme Value Modeling and Risk Analysis: Methods and Applications pp 179–194

Richards J, Wadsworth JL (2021) Spatial deformation for nonstationary extremal dependence. Environmetrics 32(5):e2671. https://doi.org/10.1002/env.2671

Saunders K, Stephenson A, Karoly D (2021) A regionalisation approach for rainfall based on extremal dependence. Extremes 24(2):215–240. https://doi.org/10.1007/s10687-020-00395-y

Schlather M (2002) Models for stationary max-stable random fields. Extremes 5(1):33–44. https://doi.org/10.1023/A:1020977924878

Schlather M, Tawn JA (2003) A dependence measure for multivariate and spatial extreme values: Properties and inference. Biometrika 90(1):139–156. https://doi.org/10.1093/biomet/90.1.139

Schütze H, Manning CD, Raghavan P (2008) Introduction to information retrieval, vol 39. Cambridge University Press Cambridge

Shi J, Malik J (2000) Normalized cuts and image segmentation. IEEE Transactions on pattern analysis and machine intelligence 22(8):888–905. https://doi.org/10.1109/34.868688

Smith RL (1990) Max-stable processes and spatial extremes. Unpublished manuscript 205:1–32

Von Luxburg U (2007) A tutorial on spectral clustering. Statistics and computing 17(4):395–416. https://doi.org/10.1007/s11222-007-9033-z

Wierzchoń ST, Kłopotek MA (2018) Modern algorithms of cluster analysis, vol 34. Springer